BUFFER TO BUFFER CREDIT RECOVERY FOR IN-LINE FIBRE CHANNEL CREDIT EXTENSION DEVICES

INVENTOR:

JAMES A. KUNZ

[0001] BACKGROUND

[0002] 1. Field of the Invention

[0003] The present invention relates to fibre channel systems, and more particularly, to in-line buffer to buffer credit recovery.

[0004] 2. Background of the Invention

[0005] Computer networks are used in every facet of modern life. These networks use high performance switching and data handling systems. Interconnected computers and high performance storage devices are commonly used. A switch is a network device at a node that sends and receives data across the network in units of frames. Various standards are used in these networks, for example, the Fibre Channel standard.

[0006] Fibre channel is a set of American National Standard Institute (ANSI) standards which provide a serial transmission protocol for storage and network protocols such as HIPPI, SCSI, IP, ATM and others. Fibre channel provides an input/output interface to

meet the requirements of both channel and network users.

[0007] Fibre channel supports three different topologies: point-to-point, arbitrated loop and fibre channel fabric. The point-to-point topology attaches two devices directly. The arbitrated loop topology attaches devices in a loop. The fibre channel fabric topology attaches host systems directly to a fabric, which are then connected to multiple devices. The fibre channel fabric topology allows several media types to be interconnected.

[0008] Fibre channel is a closed system that relies on multiple ports to exchange information on attributes and characteristics to determine if the ports can operate together. If the ports can work together, they define the criteria under which they communicate.

[0009] In fibre channel, a path is established between two nodes where the path's primary task is to transport data from one point to another at high speed with low latency, performing only simple error detection in hardware.

[0010] Fibre channel fabric devices include a node port or "N_Port" that manages fabric connections. The N_port establishes a connection to a fabric element (e.g., a switch) having a fabric port or F_port. Fabric elements

include the intelligence to handle routing, error detection, recovery, and similar management functions.

[0011] A fibre channel switch is a multi-port device where each port manages a simple point-to-point connection between itself and its attached system. Each port can be attached to a server, peripheral, I/O subsystem, bridge, hub, router, or even another switch. A switch receives messages from one port and automatically routes it to another port. Multiple calls or data transfers happen concurrently through the multi-port fibre channel switch.

[0012] Fibre channel switches use memory buffers to hold frames received and sent across a network. Associated with these buffers are credits, which are the number of frames that a buffer can hold per fabric port.

[0013] In Fibre Channel, buffer to buffer credit mechanism is used to control frame flow on a Fibre Channel link to prevent the inability to deliver any frames because of lost R_RDYs or lost frames. Fibre Channel point-to-point links lose credit when an R_RDY or an SOFx (Start Of Frame) is corrupted in transit. As credit is lost, performance degrades until frame timeouts occur. Then the only recourse is to reset the link.

[0014] The Fibre Channel standard has a credit recovery mechanism for lost R_RDYs or lost frames. Both ports on the link must support the Fibre Channel credit recovery before it can be enabled.

[0015] Fibre Channel credit recovery is used for point to point links (including links from end-user devices to switches).

[0016] Fibre Channel credit recovery defines a BB_SC_N number from 0 to 15 and two primitive signals, BB_SCr and BB_SCs. When BB_SC_N is not zero, credit recovery is enabled. Two credit recovery operations are used, one for lost frame(s) and another for lost R_RDY(s).

[0017] For lost frame(s) credit recovery, BB_SCs is transmitted whenever 2**BB_SC_N frames have been transmitted since the last BB_SCs was transmitted. The receiving port counts the number of frames received between BB_SCs primitive signals received and if the number is less than 2*BB_SC_N, it transmits as many R_RDYs as frames were lost back to the originator of the frames. Thus the originator does not lose credit for transmitting more frames.

[0018] For lost R_RDY(s) credit recovery, BB_SCr is transmitted whenever 2**BB_SC_N R_RDYs have been transmitted since the last BB_SCr was transmitted. The receiving port counts the number of R_RDYs received

between BB_SCr primitive signals received and if the number is less than 2*BB_SC_N, it adds the lost number of credits to its credit counter. Thus the receiver does not lose credit for transmitting more frames.

[0019] Most FC switches have approximately 8-323 credits per fabric, which meets the requirements for shortwave links. However, the demand for longer links is increasing as networks are being spread globally.

[0020] A fibre channel credit extender is very desirable for long range communication. Typically, this extender is placed between an end node and an optical repeater.

[0021] Conventional credit extenders do not accurately maintain buffer to buffer recovery information between its input and output interface as described above. This can result in disruption of network traffic.

[0022] Therefore, what is required is a process and system that efficiently maintains buffer to buffer recovery information in credit extenders.

[0023] SUMMARY OF THE INVENTION

[0024] In one aspect of the present invention, a method for credit recovery of lost frames in an in-line credit extender coupled between a remote device and a local device is provided. The method includes, comparing received frame count and a first programmed counter value when BB-SCs are received; loading the difference

between the programmed counter value and the received frame count into a buffer and to a first counter that counts each frame that is transmitted; and sending BB-SCs to the local device if there is a match between the first counter value and a second programmed counter value.

[0025] The first and the second programmed counter values are the same. The number of buffer credits lost are determined by the difference between the first or second programmed counter value and the received frame count.

[0026] In yet another aspect of the present invention, a system for credit recovery of lost frames in an in-line credit extender coupled between a remote device and a local device is provided. The system includes, a first counter for counting received frames; a first programmable counter that is programmed with a value; a comparartor for comparing the first counter and the first programmable counter value when BB_SCs are received; and a second counter for counting transmitted frames.

[0027] The system also includes, a second programmable counter whose value is compared to the second counter and if there is a match between the two values, BB-SCs are sent to the local device.

[0028] In yet another aspect, a method for credit recovery of lost R_RDYs in an in-line credit extender coupled between a remote device and a local device is provided. The system includes, counting received R_RDYs, wherein a first counter counts the received R_RDYs; setting a flag when a BB_SCr is received; and transmitting BB-SCr when the first counter value is zero and the flag is set.

[0029] The method also includes, counting R_RDYs after BB_SCrs are received, wherein a second counter counts the R_RDYs; and transmitting R_RDYs when the second counter value is non-zero.

[0030] The first counter value is decreased everytime an R_RDY is transmitted and the flag is cleared after a BB_SCr is transmitted. Also, the second counter is decremented everytime an R_RDY is transmitted.

[0031] In yet another aspect of the present invention, a system for credit recovery of lost R_RDYs in an in-line credit extender coupled between a remote device and a local device is provided. The system includes, a first counter for counting received R_RDYs; a second counter for counting R_RDYs received after BB_SCrs are received; and a R_RDY control module that transmits R_RDYs when the first counter value is non-zero.

[0032] The system also includes, a register that sets a flag when a BB_SCr is received; and a BB-SCr control module that transmits BB_SCrs when the first or second counter value is zero.

[0033] In one aspect of the present invention, the local device is in sync with the remote device, and credit management occurs efficiently.

[0034] This brief summary has been provided so that the nature of the invention may be understood quickly. A more complete understanding of the invention can be obtained by reference to the following detailed description of the preferred embodiments thereof concerning the attached drawings.

[0035] BRIEF DESCRIPTION OF THE DRAWINGS

[0036] The foregoing features and other features of the present invention will now be described with reference to the drawings of a preferred embodiment. In the drawings, the same components have the same reference numerals. The illustrated embodiment is intended to illustrate, but not to limit the invention. The drawings include the following Figures:

[0037] Figure 1 is a block diagram of a fibre channel network system;

[0038] Figure 2 is a block diagram showing a fibre channel extender between a local device and a remote

device, according to one aspect of the present invention;

[0039] Figure 3 is a system diagram showing how BB_SCs are handled, according to one aspect of the present invention;

[0040] Figure 4 is a block diagram showing how BB_SCrs are handled, according to one aspect of the present invention;

[0041] Figure 5 is a flow diagram for handling BB_SCs, according to one aspect of the present invention; and

[0042] Figure 6 is a flow diagram for handling BB_SCrs, according to one aspect of the present invention.

[0043] DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0044] Definitions:

[0045] The following definitions are provided as they are typically (but not exclusively) used in the fibre channel environment, implementing the various adaptive aspects of the present invention.

[0046] "BB_SCs": Flow control primitive signal used for credit recovery involving lost frames.

[0047] "BB_SCr": Flow control primitive signal used for credit recovery of lost R_RDYs.

[0048] "E-Port": A fabric expansion port that attaches to another Interconnect port to create an Inter-Switch Link.

[0049] "F-Port": A port to which non-loop N_Ports are attached to a fabric.

[0050] "Fibre channel ANSI Standard": The standard describes the physical interface, transmission and signaling protocol of a high performance serial link for support of other high level protocols associated with IPI, SCSI, IP, ATM and others.

[0051] "FC-1": Fibre channel transmission protocol, which includes serial encoding, decoding and error control.

[0052] "FC-2": Fibre channel signaling protocol that includes frame structure and byte sequences.

[0053] "FC-3": Defines a set of fibre channel services that are common across plural ports of a node.

[0054] "FC-4": Provides mapping between lower levels of fibre channel, IPI and SCSI command sets, HIPPI data framing, IP and other upper level protocols.

[0055] "Fabric": The structure or organization of a group of switches, target and host devices (NL_Port, N_ports etc.).

[0056] "Inter-Switch Link": A Link directly connecting the E_port of one switch to the E_port of another switch.

[0057] Port: A general reference to N. Sub.-- Port or F.Sub.--Port.

[0058] "N-Port": A direct fabric attached port.

[0059] "R_RDY": Flow control primitive signal used for establishing credit. Receiving an R_RDY increases credit, while sending an R_RDY decreases credit.

[0060] "Switch": A fabric element conforming to the Fibre Channel Switch standards.

[0061] To facilitate an understanding of the preferred embodiment, the general architecture and operation of a fibre channel system will be described. The specific architecture and operation of the preferred embodiment will then be described with reference to the general architecture of the fibre channel system.

[0062] Figure 1 is a block diagram of a fibre channel system 100 implementing the methods and systems in accordance with the adaptive aspects of the present invention. System 100 includes plural devices that are interconnected. Each device includes one or more ports, classified as node ports (N_Ports), fabric ports (F_Ports), and expansion ports (E_Ports). Node ports may be located in a node device, e.g. server 103, disk array 105 and storage device 104. Fabric ports are located in fabric devices such as switch 101 and 102. Arbitrated loop 106 may be operationally coupled to switch 101 using arbitrated loop ports.

[0063] The devices of Figure 1 are operationally coupled via "links" or "paths". A path may be established between two N_ports, e.g. between server 103 and storage 104. A packet-switched path may be established using multiple links, e.g. an N-Port in server 103 may establish a path with disk array 105 through switch 102.

[0064] Figure 2 shows a block diagram of a system 200, according to the present invention, using the various adaptive aspects of the present invention. A remote device 201 sends data via an optical converter 203 to credit extender 200A. A clock recovery/deserialzer module 207 processes incoming data. De-serialized data 210 is then sent to a receive (Rx) link engine 208 and then sent to a frame buffer 209.

[0065] Credit extender 200A is also coupled to a local device 221. This may be a host bus adapter or a switch port. When data has to be sent by credit extender 200A, it is serialized by serializer 219 and sent to the Rx buffer 223 at device 221.

[0066] Remote device 201, credit extender 200A and local device 221 must be in sync when it comes to buffer credit management. Sometimes, credit is lost in long-range communication between devices. Credit can be lost on both receive and transmit sides, i.e., between

remote device 201 and credit extender 200A, and between local device 221 and credit extender 200A.

[0067] In one aspect of the present invention, the buffer to buffer credit recovery mechanism uses BB_SCs and BB_SCr primitive signals to recover lost credit. BB_SCs are sent by a remote device after certain number of frames have been transmitted. BB_SCr is a primitive that is sent out after a certain number of R_RDYs have been transmitted.

[0068] Figures 3 and 4 describe how the receive and transmit side operate in managing BB_SCs and BB_SCrs, according to one aspect of the present invention. The systems shown in Figures 3 and 4 are located in buffer to buffer credit recovery module 216.

[0069] As stated above, BB_SCs are sent periodically by remote device 201 to local device 221 via credit extender 200A. In one aspect of the present invention, this allows the local device 221 and remote device 201 to be in sync.

[0070] Figure 3 shows a block diagram of system 300 that handles BB_SCs. BB_SCs 302 and SOF_rcd 303 (Start of Frame)are received from device 201. BB_SCs are received by module 304 that includes a counter 305 that can be programmed/hardcoded by firmware. SOF_rcd 303 is counted by counter 306. Values 305A and 306A from

counters 305 and 306, respectively, are compared when BB_SCs are received. If 305A and 306A match, then the difference 309 is zero. If the values do not match, then the difference 309 is sent to FIFO 310 (lost frame count FIFO).

[0071] Value 310A is added to counter 313 that counts the SOF of transmitted frames provided by 301.

[0072] Counter 313 output value 313A is compared to counter 314's output 314A. If there is a match, then BB_SCs 316 are sent out to local device 221. If there is no match, BB_SCs are not sent out to local device 221.

[0073] It is noteworthy that counters 314 and 305 are similar and set to the same value.

[0074] The following summarizes the Figure 3 system operation:

[0075] Receive Side:

[0076] If BB_SC_NUM 305 is set to a non-zero value, perform the following:

[0077] After receiving each frame, increment RX_BB_FRM_CNT register 306 by one. If RX_BB_FRM_CNT equals 2^BB_SC_NUM, set RX_BB_FRM_CNT 306 to zero.

[0078] When BB_SCs primitive is received, the number of BB_Credits lost is calculated using the following:

BB_Credits lost = (2^BB_SC_NUM - RX_BB_FRM_CNT) mod 2^BB_SC_NUM

For each BB_Credit lost, increment the TX_BB_FRM_CNT register 313 by one.

Thereafter:

Set RX_BB_FRM_CNT 306 to zero.

[0079] Transmit Side:

[0080] If BB_SC_NUM 314 is set to a non-zero value, perform the following:

After transmiting each frame, increment TX_BB_FRM_CNT register 313 by one.

Send BB_SCs primitive 316 if TX_BB_FRM_CNT 313A equals 2^BB_SC_NUM 314A.

If TX_BB_FRM_CNT 313 equals 2^BB_SC_NUM 314, set TX_BB_FRM_CNT 313 to zero.

[0081] Figure 4 shows system 400 for handling BB_SCrs, according to one aspect of the present invention. R_RDYs 401 are received from remote device 201 and are counted by counter 410 before BB_SCr 403 is received. When BB_SCr 403 is received, a BB_SCr flag is set in register 403 (also referred to Bb_SC_RDY register), which indicates that a BB_SCr needs to be sent to local device 221. The following describes how BB_SCrs are handled, according to one aspect of the present invention.

[0082] (a) R_RDY 401 are received and counted by counter 410 and when counter 410 value 413 is non-zero, an R_RDY 419 is sent out by R_RDY control module 420. After R_DY 420 is sent out, counter 410 is decremented by one.

[0083] (b) When BB_SCr 402 is received, BB_SCr flag is set in register 403, which indicates that a BB_SCr needs to be sent. BB_SCr 418 are sent out when counter 410 value is zero, as shown by signal 422, thereafter, counters 410 and 409 are flipped.

[0084] (c) If an R_RDY 401 is received after BB_SCr 403 is received, then counter 409 counts R_RDYs, i.e the counting of R_RDY flips from counter 410 to counter 409. R_RDY 420 is sent when 414 is non-zero and the process continues.

[0085] The following is a summary of the foregoing steps:

(i) After receiving each R_RDY 401,

if BB_SC_RDY 403 is clear, increment PRE_BB_RDY_CNT register 410 by one.

if BB_SC_RDY 403 is set, increment POST_BB_RDY_CNT register 409 by one.

(ii) When a BB_SCr primitive 402 is received, set the BB_SCr flag in register 403 .

The register 403 flag is cleared (signal 416) when a BB_SCr 418 is transmitted.

(iii) Send BB_SCr primitive 418 if PRE_BB_RDY_CNT 410 is zero and BB_SC_RDY flag (register 403) is set.

(iv) Clear the BB_SC_RDY register 403 flag.

(v) Change the state of the BB_RDY_CNT counters (i.e. (flip counter 410 and 409).

(vi) Send R_RDY 419 if PRE_BB_RDY_CNT is non zero (i.e. value 413 is non zero).

[0100] Figure 5 is a flow diagram of executable process steps for handling BB_SCs. It is noteworthy that steps S500-S502 and S503-S505 occur simultaneously.

[0101] Turning in detail to Figure 5, in step S500, BB_SCs 302 are received from remote device 201 with frame count 303.

[0102] In step S501, counter 305 and 306 values are compared. In step S502, the difference between values 305A and 306A is loaded into FIFO 310.

[0103] In step S503, the SOF count on the transmit side 301 is received.

[0104] In step S504, transmit counter 313 is incremented after each frame transmission. The difference between values 305A and 306A is sent to counter 313.

[0105] In step S505, transmit counter value is compared with counter 314 value. If there is a match, then in step S506, BB_SCs 316 are sent to local device 221.

[0106] Figure 6 is a flow diagram of executable process steps for handling BB_SCrs, according to one aspect of the present invention.

[0107] In step S600, R_RDYs 401 are received and counted by counter 410.

[0108] In step S601, R_RDYs 419 are transmitted by R_RDY control module when counter 410 value 413 is non-zero. When an R_RDY is transmitted, counter 410 is decremenetd (see signal (or command) 411).

[0109] In step S602, BB_SCrs 402 are received. A flag is set at register 403 that indicates that a BB_SCr needs to be sent out.

[0110] In step S603, BB_SCr 418 is sent out by BB_SCr control module 417 when counter 410 is zero and the flag in register 403 is set. Register 403 is cleared when BB_SCr is sent (see signal 416)

[0111] In step S604, R_RDYs 401 are received after BB_SCrs are received. R_RDYs 401 are now counted by counter 409.

[0112] In step S605, R_RDYs 419 are sent when counter 409 value 414 is non-zero. When an R_RDY is sent out, counter 409 is decremenetd (see signal 415).

[0113] In one aspect of the present invention, the local device is in sync with the remote device, and credit management occurs efficiently.

[0114] Although the present invention has been described with reference to specific embodiments, these embodiments are illustrative only and not limiting. Many other applications and embodiments of the present invention will be apparent in light of this disclosure and the following claims.